Linked Open Data for Filmarchives - Organised by the LOD-Task Force of the FIAF Cataloging and Documentation Commission (CDC)

Date: 14th/15th of February 2019
Place: Central and Regional Library Berlin (Zentral- und Landesbibliothek Berlin, ZLB, Amerika Gedenkbibliothek)
Organised by: Adelheid Heftberger, Georg Eckes and Anna Bohn

1. Participants

Edward Anderson (British Film Institute, London)
Anna Bohn (Central and Regional Library Berlin - ZLB, Berlin): anna.bohn@zlb.de; anna.bohn.film@gmail.com
Peter Bubestinger-Steindl (Software developer, AV-RD): p.bubestinger@av-rd.com
Raymond Drewry (Chief technologist at Movielabs/EIDR): rdrewry@movielabs.com
Georg Eckes (German Federal Archive, Berlin)
Adelheid Heftberger (FIAF CDC, German Federal Archive, Berlin) a.heftberger@bundesarchiv.de; adelheidh@gmail.com
Jürgen Keiper (Stiftung Deutsche Kinemathek, Berlin): jkeiper@deutsche-kinemathek.de
Michael Loebenstein (FIAF EC, Austrian Film Museum, Vienna)
Margret Plank (German National Library of Science and Technology - TIB, Hannover): Margret.Plank@tib.eu
Mats Skärstrand (Swedish Film Institute, Stockholm)
Jakob Voß (GBV Common Library Network, Göttingen): jakob.voss@gbv.de

2. Scope and Structure of Workshop

There is a growing interest in current developments in the semantic web and linked open data (LOD) within the cultural heritage sector. It is the CDC's impression that film archives and (film) libraries are aware of this process and try to define a possible application of LOD in our field. The CDC LOD-Task Force was founded with the specific aim to address these questions and to look into possible steps in order to support film archives.

It is of interest to discuss options, necessary actions and, most importantly, the most useful infrastructure for film archives (e.g. an ontology for audiovisual media which is adapted to the requirements of film heritage institutions), it seems advisable to discuss options in a workshop dedicated to this topic. The workshop brought together experts in the field of film archiving, information specialists and computer scientists.

We asked ourselves for example: What are useful linked open data for film archives/our users? What projects/tasks do we want to do as film archives?

The organisers asked the participants to prepare a 5 minute presentation on the following:

- Recent activities carried out in your institution connected to LOD (if applicable)
- Visions, plans, possibilities
- Problems, obstacles, requests

Afterwards, we had a discussion to come up with the most important issues and questions to be discussed.
We split up in three workgroups:

- Group 1: strategic/political questions
- Group 2: practical questions
- Group 3: Focus on broader and narrower community and users

3. Summary of Input by Participants (Presentations and Discussion)

The participants spoke about available reliable resources for audiovisual material and how film archives, libraries and other heritage institutions can link to those resources. It was agreed upon that we need stable and persistent identifiers (e.g. national filmographies, Wikidata, EIDR, ISAN etc.) and must find ways of how to deal with wrong filmographic data.

We discussed the role of the FIAF CDC as a group which can provide information on the topics above, e.g. research (linked open) data sources and also find out which LOD resources needed by the FIAF community are not available yet, maybe conduct a survey on the resources and practices in place, and possibly give recommendations. The CDC LOD Task Force could also tackle mapping entities to the EN 15907 schema, although some pioneering work has been done already by the Cinematheque Royale de Belgique.

It was brought up that starting using LOD means a change in the institutions policy. Using LOD instead of trusting only one's own metadata means opening up the collection to the external world to some degree. There is a growing sense that it is useful to pool resources when it comes to cataloging and thus being able to make use of the expertise of specialised knowledge. Using LOD also means starting to think about data governance and the governance of ontologies.

There are pitfalls to be aware of, like how to deal with coexisting and uncontrolled ontologies/vocabularies, e.g. ontology mapping. Using or even writing ontologies also means having to face the nuances and differences in languages which can not easily be solved. It was emphasised that the question whether ontologies are not too limited when describing reality should not be taken too lightly. We were made aware of the MovieLabs ontology (https://movielabs.com/creative-works-ontology/).

We discussed how film archives can not only use available LOD but how archives could contribute to existing databases like Wikidata, EIDR etc.

There were technical practicalities mentioned by the participants: If an archive wanted to use LOD, it is actually not easy to get RDF triples. Also how to search (SPARQL vs. GraphQL) resources, how to access the data (e.g. via APIs and/or data dumps?), which output format to use (JSON? other?), how to reference data sets (DOI, CrossRef, etc)? Participants report of experiments with machine learning (e.g. genres and catalog matching) and using Wikidata within their film archives. One problem encountered already when trying to share data via API were the new GDPR (General Data Protection Regulations).

We were posing questions of how to enable search over different collections and media (multi modal search) and how to connect not only our collections but also reach out to researchers and the public for certain topics (subject indexing)!

We discussed possibilities of combining audiovisual analysis (e.g. machine learning, computer vision) with indexing (how granular do we want to get) and tracking reuse of our

digital material automatically (analog to text citation). How can LOD help to mine our own databases?
We discussed the idea of not only limiting automatic validation to filmographic records but also extend it to manifestation data like film lengths (e.g. is my item identical with other items from other archives?)

One participant stated clearly that in his opinion ontologies are overrated, but the important part are the authority files and identifiers. There is a mapping project to facilitate mapping between different authority files (https://coli-conc.gbv.de/).

We talked about priorities for LOD, and agreed that film archives need to think about what is most important for them. Possible starting points are filmographic records and authority records, but also subject indexing, controlled vocabularies on item level and information about our collections (provenance data) could be tackled.

There was an agreement that using LOD will help overcome shortage of manual resources in cataloging. Film archives should invest in resources and not in unnecessary double (redundant) work, especially when it comes to filmographic data and authority records. Importing data, and/or collaboratively improving common data sources, instead of entering data yourself (when possible), is one logical step in the right direction.

While some participants suggested to strive towards a union catalogue of filmographic data first (maybe on national level) and then start enriching the collection(s) with external sources, others were in favour of breaking it down into smaller (decentralised) tasks, working on different (pilot) projects creating vocabularies (e.g. value lists for credits and casts).

We need to think about how we want to use LOD: e.g. just to copy & paste links into one's own database vs. using scripts to import/export data from external sources. We need to set up technical processes within film archives in order to implement automated processes. Ideally we can even build recommender systems for linking external data and work via interfaces. Decisions like these have great impact on how cataloging is performed in the future but also on budget plans and required skills of current/future employees.

We agreed that in order to start using (or creating) LOD within film archives we need to also talk about training courses for our staff.

4. Workgroups

a) Results Workgroup 1 on strategic/political questions

In this workgroup we tried to come up with a definition of what LOD actually means for us in this workshop. Furthermore we collected convincing arguments for people within film archives of why they should look into LOD and in what way these actions are beneficial to them.

For the participants of the workshop LOD is not so much a fixed concept, but can be rather implemented taking a scalable approach when wanting to start integrating LOD in film archives. The scale and steps that can or should be taken depend on the size of the institution, human and financial resources and other factors.

Generally times seem good to start thinking about new concepts because many film archives are changing their databases and implement data models like CEN 15907. We also discussed some advantages of an union catalogue. Such a catalogue may reduce cataloguing efforts and - more important - help to identify or map identical manifestations & prints. Nevertheless a broader concept (including semi automatic tools) would be necessary.

Before engaging in LOD it is advisable to work on a cataloging policy and handbook in order to establish cataloging practises and define mandatory fields (see CEN 15907 and FIAF Cataloging Manual for reference). It is advisable to transform our databases into machine readable formats, allowing for operations both from the human cataloger and automated processes.

There will always be a certain amount of intellectual interference necessary, but we need to build/use appropriate tools for disambiguation and film identification. We need to address the human factor and the fears/needs of our staff. As an institution we can define our stakeholders and then develop a *minimal-LOD-version* as well as a *maximum-LOD-version* and then decide what is best and what can be achieved.

Every institution is different, thus we have to decide how to deal with dependence on online services (e.g. are my links updated live) and possible collaboration with data models for union catalogs. Can I/do I want to be part of a pilot project or join later.

b) Results Workgroup 2 on practical questions

Workgroup 2 discussed the following three questions:

1. Which resources are trustworthy?

It was stated that it is crucial to know which sources are trustworthy at what *kind of information* in a some *context* (e.g. limited to some years, areas...). For instance Wikidata is good at maintaining unique identifier mappings but Wikidata is somewhat biased towards English language titles, regardless of a work's original release context. For instance IMDb gives minimal context on release dates (e.g. country, type as festival, streaming, DVD...).
There are some authoritative sources on release dates of films for selected areas.
Can "trusty" facets or metrics be expressed alongside data to qualify its trustworthiness?
It is also crucial to know which sources use what kind of model.
The workgroup said, that the two best sources of unique identifiers and linked identifiers are Wikidata and EIDR, both are recommended. In some national contexts ISAN to a certain extent may have to be included, due to existing cooperations between national film archives and ISAN agency (e.g. France).

2. What can film archives do if they want to get engaged with LOD?

You don't have to start with RDF! Connect your data with the most relevant and most rich data sources (Wikidata, EIDR, other archives that may already have open machine-readable data, ...) at least on the work-level. Use existing authority files, vocabularies, metadata schemes, ontologies... and/or map your own data to existing formats and vocabularies.

Make your own data and vocabularies (or at least parts of it) available in machine-readable format preferably under free license (CC0/public domain). The minimum data to share is **your identifiers** and additional information to disambiguate the entities. Define URIs for your identifiers. Adding some valuable information such as summaries could also be licensed under another license. Enrich your own data with external sources (with provenance); record provenance: source, type, territory.

3. Ontology: If yes, when does it make sense and for what?

Spread information about existing ontologies and vocabularies. Register information about them at http://bartoc.org (provide on FIAF CDC resources). Make vocabularies accessible in machine-readable form. Use URIs for identifiers.

Data models and vocabularies can be used independently from each other. Some agreement on what *kind of information* people and archives are interested in (e.g. original titles, release dates...) and what *contexts* information comes from would be helpful.

Recommendations of how to deal with provenance could be helpful (this is not part of EN 15907 or other standards).

Create an RDF representation of EN 15907. An XML schema already exists: http://filmstandards.org/schemas/EN15907-d1/ (provide on FIAF CDC resources). Look also at: http://filmstandards.org/schemas/EN15907-AppDefs-BA-DIF-COOP/ (provide on FIAF CDC resources). Ontologies must be published where it can be found easily, e.g. Movielabs: https://movielabs.com/creative-works-ontology/

c) Results Workgroup 3 with a focus on broader and narrower community and users

Workgroup 3 discussed possible next steps for the film archive community in terms of participation/sharing of data, and how to connect our own collections and external sources.

The following concrete suggestions/possible next steps were given:

In any case, it will probably be useful to make an RDF implementation of EN 15907. Most of the intellectual work necessary to accomplish this has already been done in the CEN standardisation committee and in the individual implementations by various institutions. What remains to be done is the actual encoding.

Also, it could be worthwhile to research whether and how the FIAF glossary of filmographic terms can be used for building the semantic relations that connect the EN 15907 entities. It could be presented as SKOS and mapped to EN 15907.

The workgroup also discussed which entities are best to start with in LOD pilot projects:

1. Film Works?
2. Persons/Names
3. Locations
4. Events
5. Items?

A practical idea which would not require much technical skills would be a "Wikidata pilot project": To make a test with entering a certain number of film titles systematically for a certain period of time and evaluate and analyse what users enter. The goal would be to measure how good the retrieved metadata is by analysing what information is added for which (kinds of) works by the Wikidata/Wikipedia community. The overall goal of such a

pilot project would be to establish what works and what does not work in terms of user engagement and crowdsourcing.

The discussion about identifiers is usually focussed on works, manifestations and persons/corporate bodies. But projects like the Joint Catalogue of Holdings (Bestandskatalog, Germany) show that attention also needs to be paid to the item. Can (or should) the problem of data about identical items in several databases be remedied by a DOI-like identifier for film items?

Automatic metadata extraction and enrichment is second compared to establishing entities and opening up metadata.

Less replication - more efficiency - more time and resources for what matters more and what can be done by the film archive ONLY.

5. Conclusions and possible next steps

The majority of the workshop participants were in favour of small steps and pilot projects to gather experience. They also identified the creation of value lists and vocabularies as next steps to take for the community. They strongly recommended using unique identifiers, internal ones but also external ones like Wikidata or EIDR identifiers. They should make their metadata “FAIR”: Findable, Accessible, Interoperable, Re-usable.

Start with:

- Film Works
- Persons/Names
- Locations
- Events
- Items

Later:

- Thesaurus for subject headings, linking to other collections
- Time-based annotation (IIIF), re-use detection/declaration (is part of), marker within documentary films for track of re-use (citation standards)

Possible next tasks for the FIAF CDC::

- start working on turning the FIAF vocabularies into LOD resources and publish them.

- look at Wikidata instances and add instances which film archives need.
- help in disseminating information about archives already using LOD by presenting their work in workshops.
- start training sessions on data literacy and getting started with LOD.
- connect more with the library and semantic web community.

Best practice examples and ideas for pilot projects are welcome. For example the *FIAF Treasures* (60.000 items) or an unrestricted subset, e.g. American Film Institute available from the LoC.

Conferences and special interest groups mentioned:

- Wikidata conference (next autumn 2019)
- IIIF Conference 2019 in Göttingen, Germany: https://iiif.io/event/2019/goettingen/
- Semantic Web in Libraries Conference (SWIB) (November 25-27, 2019 in Hamburg, Germany)
- IFLA Audiovisual and Multimedia Section: https://www.ifla.org/avms
- IFLA Linked Data Special Interest Group: https://www.ifla.org/lidasig

Written by Adelheid Heftberger, 27.3.2019